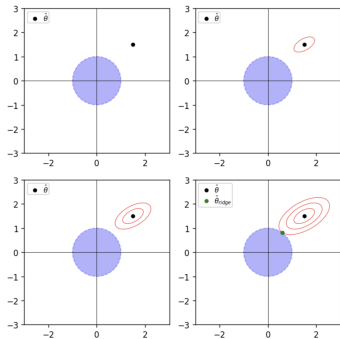


# Introduction to Machine Learning

## Regularization

## Ridge Regression



### Learning goals

- Regularized linear model
- Ridge regression /  $L_2$  penalty
- Understand parameter shrinkage
- Understand correspondence to constrained optimization

# REGULARIZATION IN LM

- Can also overfit if  $p$  large and  $n$  small(er)
- OLS estimator requires full-rank design matrix
- For highly correlated features, OLS becomes sensitive to random errors in response, results in large variance in fit
- We now add a complexity penalty to the loss:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^n \left( y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2 + \lambda \cdot J(\boldsymbol{\theta}).$$



# RIDGE REGRESSION / L2 PENALTY

Intuitive measure of model complexity is deviation from 0-origin; coeffs then have no or a weak effect. So we measure  $J(\theta)$  through a vector norm, shrinking coeffs closer to 0.



$$\begin{aligned}\hat{\theta}_{\text{ridge}} &= \arg \min_{\theta} \sum_{i=1}^n \left( y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \lambda \sum_{j=1}^p \theta_j^2 \\ &= \arg \min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_2^2\end{aligned}$$

Can still analytically solve this:

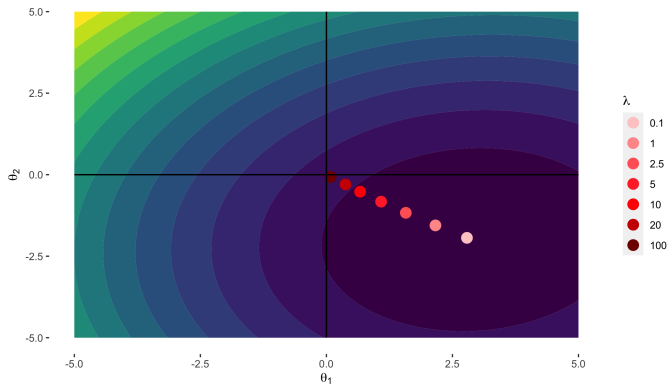
$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Name: We add pos. entries along the diagonal "ridge" of  $\mathbf{X}^T \mathbf{X}$

# RIDGE REGRESSION / L2 PENALTY

Let  $y = 3x_1 - 2x_2 + \epsilon$ ,  $\epsilon \sim N(0, 1)$ . The true minimizer is  $\theta^* = (3, -2)^T$ , with  $\hat{\theta}_{\text{ridge}} = \arg \min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda \|\theta\|^2$ .

Effect of L2 Regularization on Linear Model Solutions



With increasing regularization,  $\hat{\theta}_{\text{ridge}}$  is pulled back to the origin (contour lines show unregularized objective).

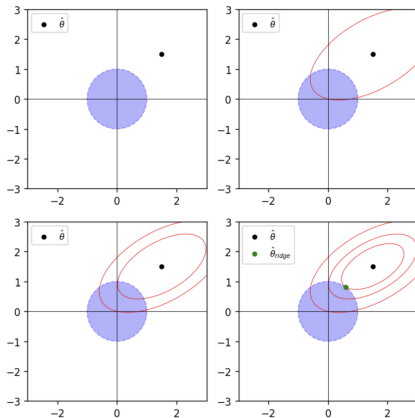


## RIDGE REGRESSION / L2 PENALTY

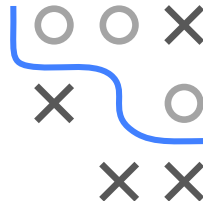
$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^n \left( y^{(i)} - f(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_2^2 \leq t \end{aligned}$$

NB: There is a bijective relationship between  $\lambda$  and  $t$ :  $\lambda \uparrow \Rightarrow t \downarrow$  and vice versa.

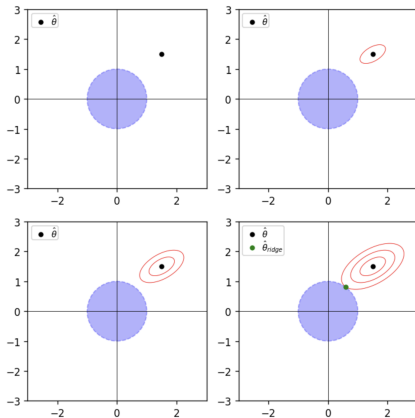
# RIDGE REGRESSION / L2 PENALTY



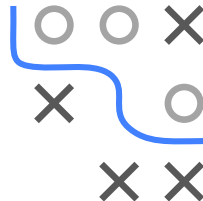
- Inside constraints perspective: From origin, jump from contour line to contour line (better) until you become infeasible, stop before.
- We still optimize the  $\mathcal{R}_{\text{emp}}(\theta)$ , but cannot leave a ball around the origin.
- $\mathcal{R}_{\text{emp}}(\theta)$  grows monotonically if we move away from  $\hat{\theta}$  (elliptic contours).
- Solution path moves from origin to border of feasible region with minimal  $L_2$  distance.



# RIDGE REGRESSION / L2 PENALTY

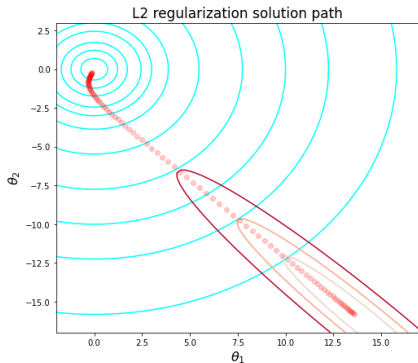


- Outside constraints perspective:  
From  $\hat{\theta}$ , jump from contour line to contour line (worse) until you become feasible, stop then.
- So our new optimum will lie on the boundary of that ball.
- Solution path moves from unregularized estimate to feasible region of regularized objective with minimal  $L_2$  distance.





## RIDGE REGRESSION / L2 PENALTY



- Here we can see entire solution path for ridge regression
- Cyan contours indicate feasible regions induced by different  $\lambda$ s
- Red contour lines indicate different levels of the unreg. objective
- Ridge solution (red points) gets pulled toward origin for increasing  $\lambda$



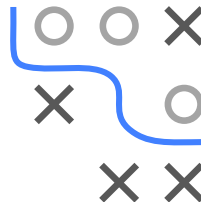
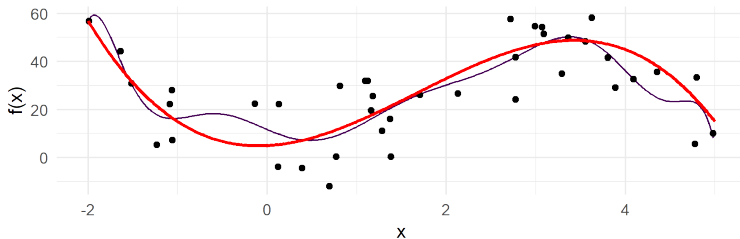
# EXAMPLE: POLYNOMIAL RIDGE REGRESSION

Consider  $y = f(x) + \epsilon$  where the true (unknown) function is  $f(x) = 5 + 2x + 10x^2 - 2x^3$  (in red).

Let's use a  $d$ th-order polynomial

$$f(x) = \theta_0 + \theta_1 x + \dots + \theta_d x^d = \sum_{j=0}^d \theta_j x^j.$$

Using model complexity  $d = 10$  overfits:



## EXAMPLE: POLYNOMIAL RIDGE REGRESSION

